From metaphors to practices

Hist. Phil. Life Sci., 33 (2011), 71-104

From Metaphor to Practices: the Introduction of "Information Engineers" into the First DNA Sequence Database[1]

Miguel García-Sancho

Departamento de Ciencia, Tecnología y Sociedad
Consejo Superior de Investigaciones Científicas (CSIC)
Calle Albasanz 26-28
28037 Madrid, Spain

ABSTRACT – This paper explores the introduction of professional systems engineers and information management practices into the first centralized DNA sequence database, developed at the European Molecular Biology Laboratory (EMBL) during the 1980s. In so doing, it complements the literature on the emergence of an information discourse after World War II and its subsequent influence in biological research. By analyzing the careers of the database creators and the computer algorithms they designed, I argue that from the mid-1960s onwards information in biology gradually shifted from a pervasive metaphor to be embodied in practices and professionals such as those incorporated at the EMBL. I then investigate the reception of these database professionals by the EMBL biological staff, which evolved from initial disregard to necessary collaboration as the relationship between DNA, genes, and proteins turned out to be more complex than expected. The trajectories of the database professionals at the EMBL suggest that the initial subject matter of the historiography of genomics should be the long-standing practices that emerged after World War II and to a large extent originated outside biomedicine and academia. Only after addressing these practices, historians may turn to their further disciplinary assemblage in fields such as bioinformatics or biotechnology.

KEYWORDS – DNA, database, cybernetics, information, systems engineering, genomics, bioinformatics

Introduction

The emergence of an *information discourse* and its role in shaping life sciences research after World War II has been a subject of debate in the history and philosophy of biology. Scholars have raised different perspectives concerning the role and utility of understanding the gene – later the DNA molecule – as an informational entity. This understanding, for some, has been a rhetorical device used by biologists to conceptualize

[1] A substantial part of the investigations reported in this paper were conducted while developing my PhD at the Centre for the History of Science, Imperial College, London, and during a short post-doctoral stay at the Centre for the History of Science, University of Manchester.

the structure and functioning of the gene (Doyle 1997; Kay 2000; Sarkar 1996a; b; 2005; Brandt 2005; Segal 2003, ch. 7). For others, DNA is literally information and its constituent sequence of nucleotides should be conceived as a system of signs, suitable to be studied from the perspective of semiotics (Hoffmeyer and Emmeche 1991; Stegmann 2005; Emmeche 1991; Hoffmeyer 1996). A different but interrelated controversy is whether this informational view of the gene has been positive for biology. There are a number of scholars who argue that information has misled biologists (Moss 2004; Griffiths 2001; Sarkar 1996b). Others, on the contrary, concede that the informational understanding of DNA has helped the conformation of disciplines, such as molecular biology, and its subsequent application to evolution and other life science problems (Kay 2000; Maynard-Smith 2000; Suárez Díaz 2007; Creager and Gaudillière 1996; Rheinberger 2006; Segal 2003, ch. 7).

The historiographical standard on the impact of informational thinking in biology has been set by, among others, Lily Kay and Sahotra Sarkar. The former has shown how the convergence of cybernetics with different lines of biological research in the late 1940s led to an understanding of the gene and its activity as information transfer. The nascent discipline of molecular biology was based on such an understanding and deployed its initial research efforts around the so-called genetic "code"; i.e., how DNA specifies the conformation of proteins (Kay 1995; 2000). For Sarkar, this informational view has been decisive for the reductionist approach which has characterized the development of molecular biology. According to this approach, the molecular structure of the DNA molecule – and mainly its sequence of nucleotides – would allow one to infer the mechanisms of gene action (Tauber and Sarkar 1992; Sarkar 1996b; 1998, 139 and sqq.). Whereas recent scholarship has minimized this reductionist agenda in the early stages of molecular biology (Morange 2008; Falk 2008), Lenny Moss and Evelyn Fox Keller have argued that DNA as information has been key for the emergence of the current concept of gene, as well as its embodiment in computing technologies during the second half of the 20th century (Moss 2004; Fox Keller 1995, ch. 3). Tim Lenoir, in line with a number of social and natural scientists, has argued that biology is currently becoming an information science and that the perception of our bodies is increasingly intertwined with computers and other information technologies (Lenoir 1999; 2002; Haraway 1997; Hayles 1999; Zweiger 2001; Hood 1992; Gilbert 1992).[2]

The literature, thus, presents the information discourse as a series of

[2] For a broader discussion of the concept of information in biology and its historical role, as well as a detailed review of the literature see García-Sancho, 2007a. On the notion of information in science more generally see Segal 2003.

concepts or metaphors that shaped positively or negatively biological research. These concepts and metaphors sometimes turned to ontologies and, due to their operative power, directed biologists, mainly in the molecular and evolutionary fields, to particular research goals (Kay 2000; Sarkar 1996b; Brandt 2005; Suárez Díaz 2007). Informational thinking has also been essential for the incorporation of information technologies such as the computer or the database into biology between the 1950s and 80s. These technologies and their effects in the development of disciplines and research agendas are becoming a main object of study in the history of biology (de Chadarevian 2002, ch. 4; Hagen 1999; 2001; Lenoir 1999; Hayles 1999; November 2004; 2006; Cook-Deegan 1994, ch. 18; Strasser 2006; Strasser 2010; Suárez Díaz 2009; 2010).

This paper contributes towards these investigations by exploring the development of the first centralized DNA-sequence database at the European Molecular Biology Laboratory (EMBL) in Heidelberg. This initiative, started in 1980, presented a key difference from previous biological database efforts: it was developed by systems and information engineers who lacked a sound biological expertise. By investigating the trajectories of these engineers, I aim to overcome the historiography which reduces the information discourse in biology to metaphors and concepts. I will argue that from the mid 1960s onwards – and especially after the late 70s – biological institutions increasingly incorporated technologies and professionals specifically designed to manage information. These technologies and professionals incorporated practices hitherto external to the life sciences.[3]

The first part of the paper will explore the background of the new EMBL database professionals and compare it to that of the creators of the first computer-based biological collections, which emerged between 1965-66. I will, then, turn to the impact of the database staff on the organization and activity of the EMBL and particularly to its difficult relationship with biological researchers during the first half of the 1980s. The paper will, finally, investigate the substantial improvement of these interactions after 1985 and the cooperation between biologists and da-

[3] My approach seeks to engage with the current shift to practice and professional identity in STS literature, in line with previous investigations (García-Sancho 2009; 2010). By addressing practices such as collecting and developing computer algorithms, Bruno Strasser and Edna Suárez Díaz have proposed lineages between current genomics and, respectively, natural history and evolutionary biology. They have also shown the difficulties that the proponents of these new practices faced when attempting to be accepted within established biological disciplines (Strasser 2006; 2010; Suárez Díaz, 2009; 2010; Suárez Díaz and Anaya Muñoz 2008). More generally, John Pickstone has argued that the history of science, technology, and medicine from the 16th century may be seen as a series of changing and interacting "ways of knowing" and "working" which allow actors to cross disciplinary boundaries (Pickstone 2000; 2007).

tabase staff in the refinement of the algorithms needed to manage the DNA sequence collection. This will facilitate a series of considerations on the necessity of long-term frameworks to address the historiography of genomics, in line with recent historical research (Suárez Díaz 2010; Ankeny 2010).

The "Systems Men" and the Proposal of a European Database

The creation of a centralized DNA sequence database in Europe was first discussed in a workshop in Schönau, a small town close to Heidelberg, the location of the central headquarters of the EMBL. In this meeting, held in 1980, other initiatives apart from the database were proposed, such as the possible involvement of the EMBL in the automation of DNA sequencing and in a large-scale sequencing project. The success and generalized acceptance of the database contrasted with the reservations toward the other initiatives. One of the reasons for the database's success was that it was proposed as a common resource for the biological community which would be run by an international institution. The EMBL, as its institutional setting, would incorporate a new category of database professionals: the systems men and information engineers.

The Workshop on Sequencing and Computing

On 5 March 1980, Ken Murray, then an invited researcher at the EMBL, proposed that a number of colleagues involved in protein and DNA-sequencing attend a workshop on "the use of the computer as an aid to sequence determination." His letter of invitation stressed the new opportunities computer technology and software opened to seek "correlations between sequences and biological features." Among the topics to be discussed, Murray referred to "the development and possible automation of methods for sequence determination," as well as the establishment of "databanks and user centres." The definite agenda for the workshop set "sequence data banks" and "sequence determination" methods as the main topics of discussion. The attendees would also debate the possible initiation of a large-scale sequencing project at the EMBL.[4]

Murray's focus on computing was not just a consequence of his personal enthusiasm. He had arrived to the EMBL a year before and was not particularly fond of this technology (Murray 2007). The push to-

[4] K. Murray (1980), Invitation letter to participants in the first Schönau workshop, in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on the Schönau meetings.

wards informatics at the EMBL was due to the previous initiatives of John Kendrew, who directed the institute from its foundation in 1974. As Soraya de Chadarevian argues, Kendrew had successfully interconnected computing and biology while at Cambridge in the 1950s and he proposed to use computers in the mathematical calculations needed to reconstruct the three-dimensional structure of molecules after their analysis through X-ray crystallography (de Chadarevian 2002, ch. 4; Mols 2007). Kendrew's appointment at the EMBL resulted in a strong commitment of the institution toward the incorporation of computing technologies to biology.

Murray had also worked in Cambridge but under the auspices of Fred Sanger, the developer of the first protein and DNA sequencing techniques. He was the first to attempt to determine DNA sequences in Sanger's laboratory during the second half of the 1960s (García-Sancho 2010, 300 and sqq). In the subsequent decade, Murray moved to the Department of Molecular Biology at Edinburgh shortly after its creation and continued to use Sanger's techniques for his research on the hepatitis B virus (Hofschneider and Murray 2001). Kendrew had persuaded Murray for a long visiting fellowship at the EMBL in order to strengthen the then emerging lines of research on DNA-sequencing. This resulted in a strong involvement of the European Laboratory in the development of sequencing technologies.

The Schönau workshop was the result of Murray's realization of the potential of the computing technologies developed at the EMBL for sequencing. Murray first sought to invite leading figures on the development of sequencing methods, such as Sanger. However, many of these figures forwarded the invitation to members of the emerging community of researchers in charge of computing instruments for sequencing. The final list of participants included, among others, Rodger Staden, the developer of the first sequencing software at Sanger's laboratory.[5]

The establishment of the definite program for the workshop followed a long correspondence between Murray and the participants. They negotiated not only the topics to be addressed, but also the ways of addressing them. The program included the automation of sequencing and the possible application of automated technologies to the genome of a complex organism as main workshop topics. However, the approach to both topics was ambivalent despite the EMBL's strengths in computing and instrument development: beside the headline "automation," the

[5] K. Murray (1980), "List of invited participants" in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Camridgeshire, UK. Folder on the Schönau meetings.

Fig. 1 - Workshops of European molecular biologists at the EMBL were common in the early 1980s, when sequencing and recombinat DNA technologies were spreading among their laboratories. Courtesy of K. Murray.

program included the subtitle "how far is this desirable and what are the limitations?" Equally, the participants would discuss the convenience of the involvement of the EMBL in a large-scale venture, such as that proposed for the bacterium *E. coli*.[6]

Both automation and sequencing of a large genome received little support during the workshop, held between 24-25 April of 1980. After its conclusion, Murray explained in a letter to the attendees that the EMBL would maintain its commitment to the development of sequencing, but while uncompromising with automation, he explicitly rejected the large-scale initiative.

[6] K. Murray (1980), "EMBL Workshop on computing and DNA sequences" in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on the Schönau meetings. The debate on the involvement in a large-scale sequencing initiative was a consequence of Project K, proposed in the late 1960s by Francis Crick as a possible project to be addressed by the future EMBL. It sought a "complete solution" of *E. coli* through a full biological analysis of this organism, which included sequencing among other potentially applicable techniques (Crick 1973). Some European scientists had expressed their reserves towards this initiative (Smith 1974).

> We are discussing the possibility of developing a totally automatic nucleotide sequenator, but before making a decision in this, we would certainly like to know what is being done elsewhere. We would welcome any information that you could pass on concerning individuals, or companies, you know with whom we might discuss this matter. We are not yet prepared to commit ourselves to the determination of a major nucleotide sequence. There are attractions and temptations, but also considerable worries in what were termed projects K and H and at present we are not prepared to embark upon either. We do expect, however, to become increasingly involved in DNA sequence determination and see little difficulty in providing useful material to feed any foreseeable work in the area of method development.[7]

Murray attributed his post-Schönau letter to the fact that "some people" inside and outside the EMBL were reluctant about automation at the time of the workshop. Molecular biologists in the early 1980s wanted "to decide themselves whether there came an A [adenine] or a T [thymine] in the sequence" rather than leaving the job to a machine (Murray 2007).[8] The attitude of these researchers contrasted with other teams which were developing automatic sequencing instruments at the same time, such as Leroy Hood's group in Caltech and the start-up biotechnology company Applied Biosystems (Chow-White and García-Sancho in press; Ramillon 2007). Regarding the large-scale project, Murray argued that it was difficult to find scientists ready to "put aside what they were doing in order to embark in such a big enterprise" (Murray 2007).

The concerns with automation and large-scale sequencing contrasted with the more favorable outcome to the database proposal at the workshop. A centralized and mainly European DNA sequence repository was an ideal project to consolidate and legitimate the EMBL six years after its foundation. Historians John Krige and Bruno Strasser have shown how the establishment of the European Laboratory was marked by problems and initial skepticism. During the mid- and late-1960s, researchers grouped in the European Molecular Biology Organization (EMBO) encountered difficulties in persuading their governments to become financially involved and there were alternative projects to the creation of a new central laboratory, such as a federation of existing ones. When in 1974 the EMBL was finally opened, its promoters were relatively successful in organizing research visits and meetings such as that of Murray and the Schönau workshop (Krige 2002; Strasser 2003). There was,

[7] K. Murray (1980), Conclusion letter to participants in the first Schönau workshop, in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on the Schönau meetings.

[8] Reluctant attitudes towards automation have been common in other fields within and outside biology and academia. David Noble has analyzed similar concerns in the metallurgic industry and Joel Hagen in the introduction of computers into taxonomy (Noble 1984; Hagen 2001).

however, the necessity of a joint European initiative and the database was seen as a crucial opportunity.

Another key factor for this success was the way in which the database was presented to the European community of molecular biologists: it was to be a service rather than a technology to be developed by its own researchers.[9] The conclusions of the workshop made clear that the EMBL database would be a centralized resource directed by a specialized staff and not led by biologists having to leave other research commitments. This specialized staff came from a remarkably different background when compared to academic molecular biologists.

Information Engineers and the Problem of Interdisciplinarity

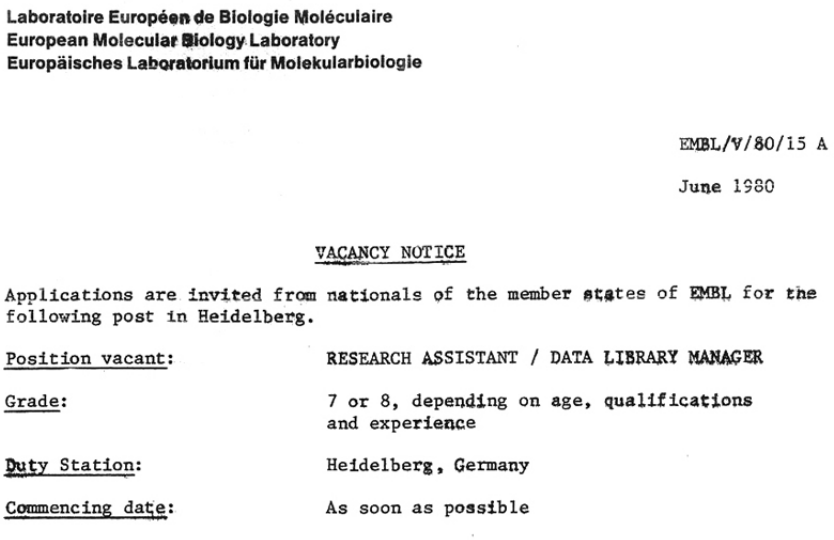
Murray's 1980 post-workshop letter stated that the EMBL had "decided to establish a nucleotide sequence data library" with the only condition of finding "the necessary staff."[10] The definition of the database as a service to the biological community in Schönau led him to seek a specific sort of professional, different from those involved in experimental biological research. A job offer attached to the letter showed that the desired staff did not have to come from inside molecular biology.

The candidates were required to have a background in "mathematics, physics or computer science" and must have previously used "numerical and statistical analysis," as well as programming languages. They should have either "made, or wish to make, the transition to research on molecular biology," but a PhD was considered a merit rather than a compulsory requirement for the job. In fact, the selected applicant would have the status of Research Assistant or Manager, instead of the higher title of Research Associate or Fellow, normally reserved for those with doctorates. Biological expertise in "problems surrounding nucleotide sequences and the structure of nucleic acids" was "desirable," but not compulsorily required.[11]

[9] The definition as an international service distinguished the EMBL database from other European repositories previously created in Germany and France. These latter collections were maintained by biologists – Richard Grantham, Heinz Schaller, and Kurt Stüber – who used the DNA sequences mainly in their own research, respectively conducted at the universities of Lyon, Heidelberg, and Cologne. Schaller was, according to Murray, the first researcher to propose centralized European DNA sequence repository and Stüber was involved in the initial stages of the EMBL database (Murray 2007; Hamm and Stüber 1982).

[10] K. Murray (1980), Conclusion letter to participants in the first Schönau workshop, in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on Schönau meetings.

[11] K. Murray (1980), "Vacancy notice" in Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on Schönau meetings.



Laboratoire Européen de Biologie Moléculaire
European Molecular Biology Laboratory
Europäisches Laboratorium für Molekularbiologie

EMBL/V/80/15 A

June 1980

VACANCY NOTICE

Applications are invited from nationals of the member states of EMBL for the following post in Heidelberg.

Position vacant:	RESEARCH ASSISTANT / DATA LIBRARY MANAGER
Grade:	7 or 8, depending on age, qualifications and experience
Duty Station:	Heidelberg, Germany
Commencing date:	As soon as possible

Fig. 2 - Job offer for the first database staff at the European Molecular Biology Laboratory (Graham Cameron's personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on the Schönau meetings. Reprinted with permission from K. Murray.

None of the professionals hired by the EMBL held a PhD, nor did any have extensive biological expertise. The first database staff, Greg Hamm, had studied both biology and engineering, and worked for many years after graduation in the emerging software industry of the United States. His role in the small company where he was based was "writing programs for various military projects, from radars to missiles." Hamm arrived in Heidelberg during the mid-1970s as the result of a leisure trip. He, then, decided to extend his stay and found a part-time job at the EMBL in sequencing software. With it, he "just wanted to make some money," rather than pursuing a research career. When the database vacancy arose, he applied and was surprised at his success, since "other candidates seemed stronger, with PhDs in biology" (Hamm 2007).

Graham Cameron, hired in 1982 to help Hamm, had abandoned a degree in psychology and worked in maintaining a database with household information at the University of Essex (UK). He was familiar with database technology, which at that time was rather alien to biological research. Cameron defined himself as an "information engineer," specialized in dealing with and organizing large amounts of data, but with no expertise in biology or academic science (Cameron 2007).

Both Hamm and Cameron agree that biological skills were not essential during their early years at the EMBL. For the latter, the crucial database problem was "understanding information," and it was secondary whether this information belonged to biology or to another realm (Cameron 2007). For Hamm, managing a database was a matter of engineering systems, similar to the ones he had worked with in the computing industry and did not require very sophisticated expertise in biology.

> I was probably the only one who was looking at [the database] as an engineering task rather than as a scientific task. Obviously, there was an important scientific content, it was necessary to understand the science, but basically I thought it as an engineering task in which the problem was how to collect, edit, curate and distribute the body of scientific data around DNA sequencing. And I think this didn't require any new discovery about how nature worked; it required an awful lot of systematic work around handling and refining data. (Hamm 2007)

Hamm and Cameron belonged to an engineering tradition which had experienced a considerable expansion during the 1940s and 50s, mainly outside academia. Historian David Mindell has called this tradition "systems sciences" and defined it as an approach derived from different "engineering cultures" which emerged in the late 19th century and "coalesced as [World War II] ended" (Mindell 2002, 8). These cultures commonly conceptualized the interactions between man and machine as a system marked by exchanges of information. With these information exchanges, the operator sought to lead the device towards a desired response. The military anti-aircraft batteries, the early computers, and the telephone were all, according to Mindell, inspired by the same principle. Each culture developed independently until World War II and triggered multiple post-war applications under the common umbrella of systems sciences (Mindell 2002, 7-11).

Hamm's defense programming was one of those post-war applications. By using the computer, he analyzed the variables involved in the movement of a missile or radar target – weather conditions, shape, speed, mechanical attributes – with the aim of predicting its behavior. This way, the missile or radar operators could use the computer to determine the future positions of their target in order to either shoot or trace it. This system analysis of interrelated variables, according to Hamm, could also be used to handle DNA sequences, even if he was not an expert in biology (Hamm 2007).

The application of systems sciences to biology was not a novelty in the early 1980s, at the time of Hamm's migration to the EMBL. Between the late 40s and 50s, as Lily Kay has shown, an increasing number of biologists adopted key notions of these sciences, such as control, feedback or

information, in order to investigate gene action.[12] They used two widespread branches of systems sciences, Norbert Wiener's cybernetics and Claude Shannon's mathematical theory of communication, to analyze the so-called "coding problem"; i.e., the specification of proteins by genes. By mathematically quantifying the exchange of information between genes and proteins, these biologists aimed to predict which proteins a gene determined without the necessity of analyzing them experimentally (Kay 2000, chs. 3-4; García-Sancho 2007a, 17 and sqq.; Segal 2003, Part I and ch. 7; Hayles 1999). During the late 50s and 60s, François Jacob and Jacques Monod used the concept of feedback in combination with biological experiments, in order to model the regulation of genes by enzymes (Creager and Gaudillière 1996; Kay 2000, ch. 5; Rheinberger 2006; Fox Keller 1995, ch. 3).

The use of systems science in biology during the 1950s and early 60s was primarily based on linear models. Shannon and Wiener's communication theories presupposed a straight flow between the information source and the destination, with or without feedback. The interaction between DNA, proteins, and enzymes in gene expression and regulation perfectly suited this scheme, which informed research in both areas even after biologists shifted from mathematics to more experimental approaches (Kay 2000, chs. 4-6; Sarkar 1996b; Creager and Gaudillière 1996). Evelyn Fox Keller has noted how Jacob and Monod's investigations on gene regulation between 1959-61 marked a gradual shift from the telegraph to the computer as the technology on which biologists modeled the functioning of the organism. Biological mechanisms were beginning to be understood as processes dependent on multiple and interrelated factors rather than linear flows of information (Fox Keller 1995, ch. 3; see also Kay 2000, chs. 5 and sqq).[13]

Hamm's programming strategies were better adapted to this multi-

[12] The use of the concept of system in biology precedes World War II. Pnina Abir-Am and Donna Haraway have shown how Ludwig von Bertalanffy, the precursor of the general system theory in the social sciences, interacted during the 1930s with many of the researchers grouped in the Biotheoretical Gathering at the University of Cambridge. Bertalanffy himself worked in theoretical biology and wrote a book on this matter in 1928, translated into English by the Biotheoretical Gathering member J.H. Woodger (Abir-Am 1987; Haraway 1976). Theoretical biology as a field arose in the late 50s and partly constructed its identity as an alternative to molecular biology (Etxeberria and Umerez 2006).

[13] This transition from linear to network triggered increasing qualifications to the central dogma of molecular biology, according to which information always flows one-directionally from DNA to RNA and proteins (Antonarakis and Fantini 2006). Historians have argued for similar shifts in the brain sciences – emergence of neural networks – and in the notion of computation between the 1970s and 80s (e.g. Olazaran 1996, 643-648). Network models currently play a key role in post-genomics and biomedical research centers, which have evolved from an assembly line to a group-based organization (Burian 2007; O'Malley 2007; Hilgartner 2004; Ramillon 2007).

factor perspective. The predictive power of his programs while at the software industry was based on interlinking various sources of data rather than on a one-directional flow of information. Both in the software industry and later at the EMBL, he adopted a network instead of a linear approach to systems sciences. The database he designed was from the beginning conceived not as a model of information transfer, but as a repository to link DNA sequences with other sorts of data.

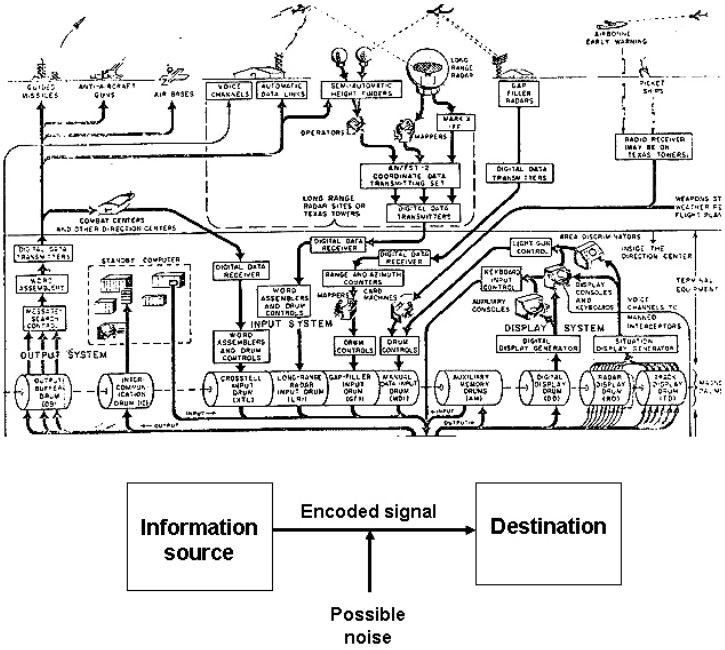


Fig. 3 - An anti-aircraft application of systems sciences (above) in contrast with Shannon's mathematical model of communication (below). The linearity of the latter contrasts with the network structure of the former (DigiBam Computer Museum, reprinted with permission, and elaborated by author from Shannon 1948).

The transition from linear to network models was favored by a number of database applications of systems sciences during the late 1950s and 60s. The owners of large companies – both civil and military – started to hire computer and systems engineers to organize the records of their multiple offices. Historian Thomas Haigh calls these new professionals "systems men" and defines their goal as transforming information management to a predictive science. They sought to create a "totally integrated management information system," which would be able to provide the company owners with "vital intelligence" about the firm (Haigh 2001, 15-16; García-Sancho 2009, 259-261). As in Hamm's pro-

gramming, the systems men aimed to make predictions from comparison of different types of data.

Information management penetrated public administration and became common in government offices, universities, and libraries during the 1960s and 70s. Cameron's label of information engineer originated within this new expertise in business and civil service (Haigh 2001, 18; Kline 2006, 528). In his previous job at the University of Essex, Cameron had become familiar with designing or, in his own words, engineering systems which allowed him to combine different household data and to derive new knowledge from such combinations. A database with information about citizens, property, and age could, for instance, determine which citizens had the largest amount of property and what was their average age.

The entrance of Hamm and Cameron into the EMBL qualifies some academic literature, which considers interdisciplinarity a distinctive feature of late 20th century biomedicine. This scholarship argues that the interaction between molecular biology and computer science fostered a number of new biomedical fields in the 1980s, most notably genomics (Lenoir 1999; 2002; Haraway 1997; Hayles 1999; Zweiger 2001; Moody 2004). However, the literature overlooks non-disciplinary professionals who, like Hamm, Cameron, or other systems men, were external to the academic world.[14]

Hamm and Cameron's non-disciplinarity illustrates that the emergence of genomics did not just follow from molecular biology, its interaction with computing, and its transformation into a discipline called bioinformatics. Genomics rather absorbed a multiplicity of pre-existing practices (Powell et al. 2007, 13 and sqq.) which originated not only in biology and other academic disciplines, but also within the business and administrative worlds. This leads to another key transformation triggered by the incorporation of Hamm and Cameron to the EMBL: the shift in the use of systems sciences in biology from modeling metaphors to practices and professionals.

Data Management Practices and Previous Biological Collections

The incorporation of systems men such as Hamm and Cameron reflected a transition in the application of informational categories to biology. If between the 1950s and early 60s information had been a "met-

[14] This neglect of non-disciplinary professionals links with the misleading identification David Edgerton has denounced between research and academic research in some STS literature, as well as with the artificial separation between basic and applied research (Edgerton 1999).

aphor" or "discourse" used as a model of gene action and regulation (Sarkar 1996b; Kay 1995; 2000; Doyle 1997; Brandt 2005; Creager and Gaudillière 1996; Rheinberger 2006; Segal 2003, ch. 7), from 1965 onwards it was increasingly embodied in practices applied to technologies, such as the database. These practices far preceded the EMBL project. Nevertheless, Hamm and Cameron were among the first professional system engineers and information managers in charge of a biological collection.

Computer-based biological databases emerged in the mid 1960s, when a number of biologists with an interest in informatics started independent projects.[15] Margaret Dayhoff, Victor McKusick, and Olga Kennard created repositories with, respectively, protein sequences, genetic diseases, and molecular structures derived from X-ray crystallography (Eck and Dayhoff 1966; McKusick 1966; Kennard 1998). The databases, according to Kennard, were founded on the "belief that the collective use of data would lead to the discovery of new knowledge" which transcended "the results of individual experiments" (Kennard 1998, 159). Kennard, Dayhoff, and McKusick were, thus, introducing into biology the practices of systems science; i.e., computer-assisted comparison of information from different sources in order to make predictions (García-Sancho 2009, 259 and sqq; on McKusick's collection see Harper 2008, ch. 7).

These researchers, however, came from a biological background and divided their time between the database and investigations in their fields. The repositories they created were partially applied to their own research and partially circulated to other biologists. Kennard, based in the Cambridge Crystallographic Data Center, made a career as a crystallographer and used the information from her database to refine molecular structures (Kennard 1998). Dayhoff, at the US National Biomedical Research Foundation, has been investigated in detail by Bruno Strasser, who describes how she constructed evolutionary phylogenetic trees by comparing protein sequences from different species (Strasser 2010).

The computers used by Dayhoff and Kennard were large mainframes located in central facilities and operated by punch cards. This created a distance between the practices of compiling and processing the data, which in the case of Kennard were compiled at her center and processed

[15] Non-computerized biological repositories were common in natural history, from the 16th century (Rosenberg 2003; Secord and Jardine 1996; Müller-Wille 2003). During the late 19th and early 20th century onwards, epidemiologists and geneticists used mechanical and electric devices to make calculations and process information which have been seldom investigated by historians. Computers were used as an aid to experiments in crystallography, physiology, and the brain sciences in the mid-20th century and inspired the early computerized databases (November 2006).

in remote computing locations. The mid-1960s witnessed the emergence of the minicomputer, slower and more limited than the mainframe, but with a suitable size for use in laboratories (Ceruzzi 2000, chs. 4-6; Campbell-Kelly and Aspray 1996, 207 and sqq.). Minicomputers coexisted with mainframes during the late 60s and 70s, but as Joseph November shows, some biological fields proved especially impermeable to the new devices. This led to the design of minicomputers such as LINC, especially oriented towards the necessities of the biomedical laboratory (November 2004; 2006, chs. 4-5).

The launch of the EMBL database coincided with an increasing incorporation of computers among biological and other scientific laboratories. During the 1970s, minicomputers gradually enhanced their power and reduced their price, and this resulted in the emergence of other devices such as the microcomputer or workstation, which were especially adapted for reduced spaces and groups of operators (Ceruzzi 2000, ch. 9; Campbell-Kelly and Aspray 1996, Part 4). The collection and computer processing of data started to be conducted at the same location, but during the 1980s researchers were not always keen on the complications of early in-house computers. This resulted in the parallel introduction of professionals such as Hamm and Cameron, with specific expertise in handling the new technologies.

Another key difference between Kennard, Dayhoff, and the EMBL database was their perception by fellow biologists and funding agencies. Kennard's collection emerged in a context marked by the Cold War and the concern that Europe was lagging behind the US and the USSR in the collection of scientific data. Her first grants were aimed to address that gap and, after the 1970s, she had to lease the database entries for her database to survive (Kennard 1998; García-Sancho 2009, 264-265). Dayhoff was initially funded by the National Institutes of Health, but given the insufficiency of the budget she had to seek additional support in the US National Airspace Agency (NASA) and Atomic Energy Commission, among other institutions. Similar to Kennard, she needed to sell the database information in the form of periodically printed volumes titled *Atlas of Protein Sequence and Structure* (Strasser 2006; 2010).

Dayhoff was never accepted as an equal by the community of experimental biologists. Her work was at that time considered theoretical and outside the realm of scientific research. Additionally, biologists did not appreciate that Dayhoff sold her data to support her project. The exchanges of information within experimental biology, Strasser claims, were founded on the free circulation of data once scientific publication had provided the authors with the necessary credit. Dayhoff's database challenged this "moral economy," since apart from not being free it

made the data available before publication (Strasser 2006, 114-118).

Hamm and Cameron's database emerged at a time in which the collection and analysis of information was better valued by experimental biologists. The spread of increasingly in-house computers from the late 1970s resulted in a socio-political embrace of data, which social scientists have called information society. Governments and their citizens, together with businessmen and scientists, saw in the control and access to information – political, economic, or techno-scientific – the main means of productivity, knowledge and welfare (Castells 1996; Webster 1997; Harvey and McMeekin 2007; García-Sancho 2009). Historian of technology Ronald Kline has claimed that these transformations contributed to the emergence of the category of "information technologies." The regulation of these technologies was the object of growing political debate during the 1980s (Kline 2006, 520 and sqq.).

The social concern with information permeated biological research, especially after the emergence of recombinant DNA and the rise of biotechnology in the 1970s. Biologists – namely in the molecular field – together with public funding bodies and private investors considered that DNA sequences and other biological data were fundamental scientific and economic assets (García-Sancho 2007b; 2007a, 29 and sqq; Kenney 1986). This led to the proliferation of computerized repositories of biological information. Shortly after the EMBL initiative, the US launched a centralized DNA sequence database, GenBank, derived from a long-term contract awarded by the National Institutes of Health (NIH) to the company Bolt, Beranek and Newman, and a group of physicists working on biological problems at Los Alamos Laboratory. Dayhoff presented a bid for this project, but was not successful (Strasser 2008; 2010).

The US and European databases, rather than individual efforts, were part of top-down initiatives coordinated by central organizations, such as the NIH and the EMBL. They were conceived as repositories of promising DNA sequences and benefited from relatively stable budgets. The professionals hired for the management of the databases made sequence submission agreements with journal editors and this allowed the free and early circulation of data, making it compatible with publication credit and other community norms of the user biologists (Strasser 2006; 2008).[16] However, the stability of the European reposi-

[16] The rise of DNA sequence and other related databases during the 1980s had an impact on biological representation. These databases and the social environment in which they emerged accentuated an ongoing transition in the life sciences from collecting things to collecting data (García-Sancho 2008, 236 and sqq.). On representation in biology see Suárez Díaz 2007; de Chadarevian and Hopwood 2004; Cambrosio et al. 2008.

tory did not prevent a number of problems, mainly derived from the relationship between the EMBL biologists and the new database staff.

The Impact of Information Management on Biology

The entrance of Hamm and Cameron into the EMBL and the rise of the DNA sequence database transformed the research conducted in this institution and particularly its investigations on the development of sequencing methods. Molecular biologists working at the European laboratory saw their identity threatened by the new database professionals and initially received them rather reluctantly. This led Hamm and Cameron to develop the first versions of the database in relative isolation and to create from the literature on computer science a series of algorithms adapted to the specificities of the stored DNA sequences. The relationship with biologists improved during the second half of the 1980s, when researchers at the EMBL and other institutions began to realize the complexities of the DNA molecules. Hamm and Cameron then established a more systematic cooperation with the EMBL biologists and regularly updated the algorithms in the face of new findings about the functioning of the stored sequences.

Hierarchy and the Identity of Sequencing

The expansion of the computer during the early 1980s did not completely stop the biases of biological researchers and institutions towards database management. Recent scholarship has shown that computer experts, once introduced into life sciences laboratories, faced "vast cultural differences" with the biological staff (Cook-Deegan 1994, 285; Moody 2004; Chow-White and García-Sancho in press; Leonelli 2010). Hamm and Cameron's first years at the EMBL were marked by similar difficulties to those experienced by Kennard and Dayhoff in the preceding decades. Molecular biologists based in the EMBL considered them as a "secretariat," exclusively serving their research necessities (Cameron 2007).[17] This perception forced Hamm and Cameron to present their project as experimental research when external funding was needed. At

[17] Some journal editors with whom Hamm and Cameron attempted to make agreements for sequence submission to the database were also initially reluctant towards the proposal. Robert Cook-Deegan has shown how the editor of *Nature*, John Maddox, resisted for a long time mandatory database submission of the sequences to be published in his journal. The reasons he gave – scarce computer facilities in many laboratories, the sequences having "nothing to do with the content" of *Nature's* papers and the journal not needing to depend on a database – point to a lack of understanding of the role of the database staff, similar to that Hamm and Cameron were experiencing at the EMBL (Cook-Deegan 1994, 288-289).

that time, grants for the development of research infrastructures were scarce and undefined.

The conflict between biological and database staff affected all the activities of the EMBL particularly on sequencing development. Hamm and Cameron's appointment as new information management professionals made the practices of gathering and storing the sequence data increasingly independent from their use. Thus, sequencing gradually became a service aimed to provide researchers with data previously collected and catalogued. In this picture, the collecting and cataloguing practices were the responsibility of staff, who were considered to be less qualified than the final users of the sequences.

This new hierarchy and shift toward service work was not so marked outside the EMBL. During the second half of the 1980s, John Sulston and Alan Coulson devoted their efforts at the Laboratory of Molecular Biology of Cambridge to the compilation of a database with map and sequence information of the DNA of the worm *C. elegans*. The data was freely available and circulated for further use by biologists in their investigations on the worm (García-Sancho 2008, 133 and sqq; Ankeny 2001; de Chadarevian 2004). Sulston and Coulson were well-respected molecular biologists and have been retrospectively considered the pioneers of revolutionary genomics (e.g. Cook-Deegan 1994, chs. 3-4). In 2002, Sulston was co-awarded the Nobel Prize in Medicine for his investigations on the worm.

At the same time, Hamm and Cameron were never considered for the Nobel Prize. The 16 papers they individually or jointly published in the scientific literature between the 1980s and 90s contrast with the over 45 written by Sulston and Coulson.[18] The marginalization Hamm and Cameron suffered by the community of molecular biologists led them to work rather independently during the initial stages of the EMBL database. In those first years, Hamm and Cameron designed a series of algorithms to interpret the stored DNA sequences almost exclusively based on their expertise in information management and systems science.

DNA, Pattern Recognition and Computers as Information Processors

Hamm and Cameron's project, formally named Nucleotide Sequence Data Library, had been preceded by important developments in database technology. From the 1950s onwards, the database gradually abandoned its military connotations and was metaphorically and loosely associated with "buckets," "hubs," or "pools" of data used by public administra-

[18] Search conducted at the US National Library of Medicine (www.pubmed.org).

tion and private offices. Between the 60s and 70s, the current identity of the database as a computer application emerged and, subsequently, databases were included in packages with software to manage the entries automatically (Haigh 2006a, 33-34). Both databases and programs were designed first by computer manufacturers and then by the emerging software divisions and companies. Information Business Machine (IBM) was one of the main early developers (Campbell-Kelly and Aspray 1996, 174-76; Campbell-Kelly 2003).

These new database management programs organized the records according to various criteria. They established different models of interactions (hierarchical, network, or relational) among the distinct types of stored data. If the data were, for instance, name of employees, payroll information, and birthdates, the software permitted a search among the entries to determine employee pay rates and age. The software was adapted to the necessities of the main customers of the computing industry, insurance companies, travel agencies, banks, libraries or job centers that wanted a high level of control over the large volume of data they produced (Campbell-Kelly 2003). During the 1970s, the relational model, which allowed unrestricted combinations between the entries, consolidated as the dominant data linkage criterion (Date 1981 [1975]).[19]

None of these models was suitable for the EMBL. Hamm and Cameron, despite being aware of the developments in systems sciences and database technology, did not include interactions among the entries in the first public releases of the Sequence Data Library, in 1982 and 1986 (Hamm and Stüber 1982; Hamm and Cameron 1986). This was because the stored DNA sequences were not adaptable to any of the available database management programs.

The programs, according to Hamm, reflected a "table view of the world," since they handled the database entries as self-contained data; e.g., books borrowed by different users or mortgages requested by clients (Hamm 2007). DNA sequences, by contrast, were "different" records, formed by continuous and large strings of already interrelated nucleotide units (Hamm 2007). The aim of the operator with these records was not only to establish connections between the sequence entries, but also to find patterns in the strings and to attribute to them certain features; e.g., the presence of a gene within the sequence. This activity was named "annotating," and became Hamm and Cameron's

[19] The relational model was invented during the 1960s by IBM programmer E.F. Codd and popularized by C.J. Date (Codd 1970; Date 1981 [1975]). It was the basis for Oracle, a successful database multinational created in 1977 in Silicon Valley and founded by Larry Ellison, an autodidactic nondisciplinary entrepreneur similar to Hamm and Cameron (Wilson 1997).

main endeavor in the early years of the European database (Cameron 2007).

During the first half of the 1980s, the EMBL team created a series of work routines "on a day-to-day basis," which were then compiled into programs that systematized the management of the stored sequences (Hamm 2007). Hamm and Cameron's strategy was to analyze the literature on computer science periodically and, especially, to evaluate a textbook written in the mid-70s by the Bell Laboratories programmers, Brian Kernighan and P.J. Plauger.

Address Procesing	A.Rudloff	1.1	11-OCT-1982
"	A.Rudloff	2.1	11-JAN-1985
Tape Distribution	A.Rudloff	1.1	12-OCT-1982
"	A.Rudloff	2.1	27-MAR-1984
Writing Data on Floppy Disks	G.Cameron	1.1	29-OCT-1982
Cover Note for Pre-release Data	GNC / AR	1.1	15-DEC-1982
Processing Pre-releases	GNC / AR	1.1	15-DEC-1982
Notes on Using CMS for Data Library Management	G.Hamm	1.1	29-JAN-1983
Author Review	KST / AR	1.1	23-JUN-1983
How to Deal with Programs and Data Submitted	GNC / GHH	1.1	5-JUL-1983

Fig. 4 - List of standard operating procedures compiled by Hamm and Cameron during the development of the database (Graham Cameron's personal archive, European Bioinformatics Istitute, Hixton, Cambridgeshire, UK. Folder on memos and reports. Reprinted with permission from G. Hamm)

The book, *Software Tools,* described a series of algorithms used "every working day" by the authors to automate a number of programming operations. One of its main features was the adoption of the UNIX operating system, developed at Bell Laboratories during the late 1960s. UNIX incorporated a rudimentary text processing program. The authors claimed that they had both "tested the programs" and "typeset the manuscript" within UNIX. They described algorithms such as search and format*,* which could be used either in a text with machine-language instructions written by a programmer or that aimed to another person written in a word-processing program (Kernighan and Plauger 1976, 5).

Hamm and Cameron, despite not using UNIX, initially typed and edited the database entries in an early text processor (Hamm 2007). This, and the reading of *Software Tools,* led them to see the potential of this technology to manage the sequence entries. Text pattern recognition algorithms had already been used by Dayhoff and other researchers in the development of protein sequencing software and sequence databases during the 1960s and 70s. At that time, the algorithms were directed to the alignment of sequences before comparison and the assemblage of the different protein fragments derived from the sequencing reactions (Suárez Díaz and Anaya Muñoz 2008; Suárez Díaz 2010; Strasser 2010).

The developers of these early pattern-recognition algorithms were mainly biologists, often unaware of and duplicating efforts with the computer and software industry (García-Sancho 2008, 84-86). Hamm and Cameron, due to their background in systems sciences, designed database management programs that went beyond the problems of sequence alignment and assemblage. Among the algorithms they imported for such programs were those of spell checking and hash-coding, the latter used to search for words within a text processor. Whereas spell checking allowed them to detect typos in the entries of the Data Library automatically, hash-coding was used to locate genes within the sequences. Given that genes are always surrounded by specific DNA sequences (initiation and termination codons), it was possible to ask the computer to search for codons within the database entry. The located points were, then, automatically annotated as the beginning and end of genes.[20]

Thus, Hamm and Cameron shifted the foundations of the algorithms from English orthography, the basis of text processors, to the available biological knowledge about DNA sequences. This adaptation allowed the database to deduce features from the stored sequences automatically. The deduced features (e.g., location of genes or proteins they coded for) were included in a particular section of the entries called Feature Table. This led the EMBL database to look different from those previously developed by the computer and software industries, and its entries to be more easily associable with those of other repositories. In 1987, after extended negotiations, the Data Library, GenBank, and a more recent Japanese database unified their formats and made their entries interchangeable. This was followed by a collaboration between the EMBL database and SWISS-PROT, a Zurich-based protein sequence repository that had become the largest of its kind. As a result, the entries of both databases became interconnected through programs which translated DNA into protein sequence (Bairoch 1991).[21]

The application of text processing algorithms in *Software Tools* and in the Data Library reflected a broader shift in the use of the computer. In their classic history of computing, William Aspray and Martin Campbell-Kelly describe how from the 1950s onwards the computer was gradually

[20] G. Cameron, G. Hamm, A. Rudloff and K. Stueber (1983), "EMBL Nucleotide Sequence Data Library User Manual" in G. Cameron personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on memos and reports.

[21] The association of interrelated databases in higher order collections – e.g. DNA and protein sequence databases in a gene expression bank and software package – has been an important development of the technology during the second half of the 1980s and 90s. This has made bio-ontologies linking the information contained in different databases an essential tool in biology and an object of STS scholarship (Leonelli 2008; 2010).

"reconstructed – mainly by computer manufacturers and business users – to be an electronic data processing machine rather than a mathematical instrument" (Campbell-Kelly and Aspray 1996, 105 and sqq). This argument has been qualified by David Mindell, who in his history of systems sciences shows that engineers, the main actors in early computer design with mathematicians, understood from the beginning that computing technology was an information system. Engineers, Mindell claims, "did not build electronic digital computers simply as calculators" and always had in mind the notion of gathering, combining and exchanging different types of data (Mindell 2002, 10).

```
ID   MMIG20    MUS.MUSCUL.IG.MOPC41; DNA; 350 BP.
XX
DT   82.01.01  (first entry)
XX
DE   First two exons in immunoglobulin light chain genes from
DE   cell line MOPC41.
XX
KW   differentiated gene; immunoglobulin.
XX
OS   Mus musculus (house mouse, souris domestique, Hausmaus)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda;
OC   Mammalia; Eutheria; Rodentia.
XX
RN   [1]   (bases 1-350)
RA   Altenburger W., Steinmetz M., Zachau H.G.;
RT   "Functional and non-functional joining in immunoglobulin light
RT   chain genes of a mouse myeloma";
RL   Nature 287:603-607(1980).
XX
FT   Key        From     To       Description
FT
FT   CDS         126     176      first exon (leader peptide)
FT   CDS         303    >350      second exon (variable part)
XX
SQ   Sequence   350 BP;  80 A; 82 C; 122 T; 66 G.
     CGTGACCAAT CCTAACTGCT TCTTAATAAT TTGCATACCC TCACTGCATC GCCTTGGGGA
     CTTCTTTATA TAACAGTCAA ACATATCCTG TGCCATTGTC ATTGCAGTCA GGACTCAGCA
     TGGACATGAG GGCTCCTGCA CAGATTTTTG GCTTCTTGTT GCTCTTGTTT CAAGGTTAAA
     ATGAAACTTA AAATTGGGAA TTTTCCACTG TTTCCAACTG TGGTTAGTGT TGACTGGCAT
     TTGGGGGATG TCCTCTTTTA TCATGCTTAT CTATGTGGAT ATTCATTATG TCTCCACTCC
     TAGGTACCAG ATGTGACATC CAGATGACCC AGTCTCCATC CTCCTTATCT
```

Fig. 5 - Sample entry of the EMBL Data Library. The DNA sequence - at the bottom - is preceded by the Feature Table (FT) (Hamm and Cameron 1986, 8. Reprinted with permission from *Nucleic Acids Research).*

The development of the Sequence Data Library shows that the shift in the nature of the data to be processed was at least as important as the putative transition toward information processing. Systems engineers, including Hamm before his incorporation to the EMBL, had already designed databases and software that gathered and combined multiple types of data between the 1950s and 70s, therefore extending the use of

the computer beyond a calculator. However, the data inputs these engineers addressed were considered self-contained variables, susceptible to cross-linkage rather than internal analysis – e.g. the variable, books, could be related to other variables, users and delays, but the texts inside the books were never analyzed.

This dominance of cross-linkage persisted despite text-editing software being available to computer programmers since the 1960s. The software was applied to machine-language texts, but never "to general-purpose office work" which at that time was in the hands of clerks or secretaries equipped with typewriters, teletypes or dictating machines. The only text computers routinely processed outside programming were "short fields such as 'title' or 'last name,'" which were subsequently shown "in the appropriate places on paychecks, invoices, and printed reports" (Haigh 2006b, 6-13). The text processor, as a general-purpose program to type and edit written documents, emerged simultaneously with the personal computer and, during the 80s, gradually became the main computing application (Bergin 2006a; b).

Hamm and Cameron's development of specific algorithms to manage DNA sequences was a consequence of this shift. Due to their background in systems sciences and information management, they adapted text-processing and pattern-recognition algorithms that were then emerging in the computer industry. These adapted algorithms allowed the Data Library not only to compare, but also to deduce meaning from the stored sequences. However, the meaning to be deduced by just adapting computer industry algorithms gradually decreased during the second half of the 1980s.

Biocomputing and Cooperation with Biologists

The relatively independent work of the database team at the EMBL encountered difficulties throughout the 1980s, as "biology became more complicated" (Hamm 2007). Discoveries such as alternative splicing, the complexities of gene regulation and the significant presence of introns in DNA sequences of evolved organisms made the simple day-to-day routines of Hamm and Cameron increasingly obsolete. As the concept of the gene became more sophisticated, the algorithms inspired in text processing – which presupposed a linear and uninterrupted nucleotide string – became inappropriate.[22]

[22] The growing sophistication, changing definitions and pervasiveness of the concept of gene have been recurrent topics in the philosophy and history of biology (Beurton, Falk and Rheinberger 2000; Fox Keller 2000; Moss 2004; Griffiths and Stotz 2004; 2006; Müller-Wille and Rheinberger 2007;

This forced the database team to respond in different ways. Towards the mid-1980s, Hamm and Cameron diversified their group and welcomed members with advanced biological backgrounds. A 1985 memo described database management as a "production line" including "clerical staff," (which conducted the literature searches for sequences) computer programmers (responsible for algorithm design) and "staff with biological knowledge" in charge of the annotation of the entries. This latter staff ranged from "students" to young biology graduates.[23]

The same 1985 memo requested molecular biologists at the EMBL to help the database staff in "reviewing" the DNA sequence entries. According to the document, the Data Library was experiencing a "considerable backlog of data" whose solution required "biological expertise" from senior researchers. The tasks in which those researchers could help were the assessment of the interest of published sequences for inclusion in the database and the determination of which data should be incorporated to the entries. Both the clerical and junior biological staff in Hamm and Cameron's team either lacked the expertise or had a background that was too "generalist" to perform such duties.[24]

Hamm and Cameron's request represented an initial merger between the previously divergent biological and database staff. A more systematized one was the creation at the EMBL of the Biocomputing Unit in 1987. This new Unit developed from a combination of the former EMBL divisions of Biological Instrumentation and Computing and Applied Mathematics, where the database team was based. It sought the promotion of hybrid backgrounds in computational biology among existing and future EMBL personnel. The Biocomputing Unit played a key role in bridging biological and computing staff and in refining Hamm and Cameron's algorithms.[25]

The emergence of this new Unit resulted in formal training and a research program on bioinformatics at the EMBL. It created an insti-

Taylor 2008). An increasing number of STS scholars has also addressed the problem of "producing meaning" from the information stored in DNA sequence – and more generally biological – databases (Fujimura and Fortun 1996; Fujimura 1999; Fortun 2008; Leonelli 2008; 2010).

[23] G. Hamm (1985), "Biological expertise for the Data Library" in Graham Cameron personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on reports and memos.

[24] G. Hamm (1985), "Biological expertise for the Data Library" in Graham Cameron personal archive, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. Folder on reports and memos. In one section of the memo, the task of reviewing an entry was compared with that of "refereeing a short paper." This suggests that a meta-literature of entries was arising among the database staff, with cross-referencing and a quality control similar to those of scientific papers.

[25] The Biocomputing Unit and combined biological and computational background it promoted evokes the concept of the "moist zone," postulated in STS scholarship to overcome the rigid dichotomy between wet and dry biology (Penders, Horstman and Vos 2008).

tutional and disciplinary framework that gradually absorbed practices developed rather informally in the preceding years. This framework not only formalized the practices, but also transformed them through a multiplicity of actors working and interacting around the database in a permanent cycle. Hamm and Cameron's team produced the database entries and these entries were used in research that led to new findings. Biologists reported the new findings to the database team and its members used them to refine both the entries and database algorithms. The database, therefore, became the "meeting ground" of previously non-interactive staff (Cook-Deegan 1994, 285).

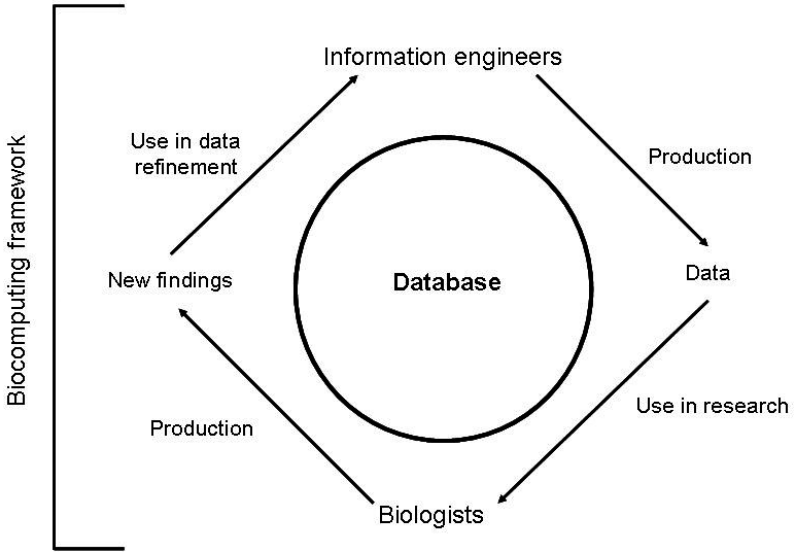


Fig. 6 - The database as a space of convergence (elaborated by author).

Conclusion: practices, disciplines and the database as a space of convergence

This paper has explored the emergence and early development of the first centralized DNA sequence database with the aim of shedding new light on the role of an information discourse in biology. My main argument is that since the mid-1960s information concepts and metaphors were increasingly embodied in data gathering and analysis practices around a number of biological databases. These practices were first performed by biologists and then by a new category of professionals in sys-

tems sciences, such as those hired by the European Molecular Biology Laboratory (EMBL) in the early 80s. The entrance of these professionals was a consequence of the incorporation of increasingly in-house computers to the laboratory and triggered organizational and disciplinary transformations in biological centers, mainly in response to their difficult relationship with biologists. This new database staff redefined the practices of gathering and analyzing biological data in light of professional computer programming and information management algorithms previously used in public administration, private companies, and offices.

The trajectories of these data-oriented practices and professionals at the EMBL have significant implications for the historiography of post-World War II biology. They show how inappropriate the history of disciplines is for analyzing the development of biological research during the second half – and especially during the last third – of the 20th century. The crucial role of systems science professionals in the European database opens the historiography of genomics and biotechnology to a multiplicity of practices, many of them outside the realm of biology and academic research. In this regard, the history of sequencing and its associated technologies cannot be solely explained by the development of academic biochemistry, molecular biology, or computer science. The first centralized DNA sequence database is a materialization of practices and professionals derived from military control and command systems, as well as from the management of other sorts of information, such as written texts, holiday bookings, or banking details.

This inappropriateness of a disciplinary framework does not mean that academic disciplines are irrelevant for the history of science. The role of the Biocomputing Unit in bridging biologists and systems scientists at the EMBL shows that regular training and an institutional line of research in bioinformatics were essential for the development of the database during the second half of the 1980s. As in molecular biology and genomics (Kay 1993; de Chadarevian 1996; 2002; Powell et al. 2007), bioinformatics acted as an umbrella and gathered a multiplicity of actors and practices which preceded its conformation. These diverse actors and practices should be the first focus of historical research for then analyzing their disciplinary assemblage.

Another factor which eases the assemblage of practices and professionals are the technologies around which they circulate. In the case of the EMBL, the database acted as a space of convergence that both gathered and embodied biological and systems science expertise. This convergence resulted in a self-replicating cycle in which biologists used the database entries in their research and their subsequent achievements were the basis for the refinement of the technology by

systems scientists. As both classical and recent scholarship has argued, biological objects and technologies occupy a boundary space and can be investigated from the perspectives of philosophy, sociology, law, and history (Star and Griesemer 1989; Gibbons et al. 2007; Leonelli 2010). Thus, biomedical databases, serve as multidisciplinary case studies for investigating cooperation and professional identities, changing scientific and social representations, and the long-term history of genomics and bioinformatics.

Acknowledgements

Special thanks are given to María Jesús Santesmases, Andrew Mendelsohn, Angela Creager and an anonymous referee for insightful comments on a draft manuscript. Other researchers involved in the development of this paper were David Edgerton and Max Stadler (Imperial College, London); Edna Suárez Díaz (Universidad Nacional Autónoma de México); Bruno Strasser (Yale University); Soraya de Chadarevian (UCLA); Adam Bostanci and other researchers at the ESRC Centre for Genomics in Society (University of Exeter); José Manuel Sánchez Ron, Javier Ordóñez, Antonio Sillero and Rafael Garesse (Universidad Autónoma de Madrid); John Pickstone, Carsten Timmermann, Duncan Wilson and other researchers at the Centre for the History of Science (University of Manchester); Richard Ashcroft (Queen Mary University, London); Hans-Jörg Rheinberger (Max Planck Institute for the History of Science, Berlin); Olga Kennard (retired); Ken Murray and Alix Fraser (University of Edinburgh); Greg Hamm (GPC-Biotech); Graham Cameron and Mark Green (European Bioinformatics Institute), Keith Benson (University of British Columbia).

Research was conducted while the author held postgraduate fellowships awarded by Caja Madrid Foundation, Madrid City Hall and Residencia de Estudiantes (Spain), as well as Imperial College, London (Hans Rausing Fellowship). Without them, it would not have been feasible. The fieldwork trips and other expenses were also covered by small grants awarded by the Royal Historical Society. The final stages in the preparation of the manuscript were supported by a Wellcome Trust postdoctoral fellowship and a contract by the Spanish National Research Council (CSIC).

References

Abir-Am P., 1987, "The Biotheoretical Gathering, trans-disciplinary Authority and the Incipient Legitimation of Molecular Biology in the 1930s: New Perspective on the Historical Sociology of Science", *History of Science*, 25: 1-70.

Ankeny R., 2001, "The Natural History of *C. elegans* Research", *Nature Review Genetics*, 2: 474-478.

Ankeny R., 2010, "Historiographic Reflections on Model Organisms: or, How the Mureaucracy May Be Limiting our Understanding of Contemporary Genetics and Genomics", *History and Philosophy of the Life Sciences*, 32(1): 91-104.

Antonarakis S. and Fantini B. (eds), 2006, "History of the Central Dogma of Molecular Biology and its Epistemological Status Today", *History and Philosophy of the Life Sciences*, special issue, 28(4).

Bairoch A., 1991, "The SWISS-PROT Protein Sequence Data Bank", *Nucleic Acids Research*, 20, Supplement: 2019-2022.

Bergin T., 2006a, "The Origins of Word Processing Software for Personal Computers: 1976-1985)", *Annals of the History of Computing*, 28(4): 32-47.

Bergin T., 2006b, "The Proliferation and Consolidation of Word Processing Software", *Annals of the History of Computing*, 28(4): 48-63.

Beurton P., Falk R. and Rheinberger H.J. (eds), 2000, *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*, Cambridge: Cambridge University Press.

Brandt C., 2005, "Genetic Code, Text, and Scripture: Metaphors and Narration in German Molecular Biology", *Science in Context*, 18(4): 629-648.

Burian R., 2007, "On MicroRNA and the Need for Exploratory Experimentation on Post-genomic Molecular Biology", *History and Philosophy of the Life Sciences*, 29: 285-312.

Cambrosio A., Jacobi D., Keating P., 2008, "Phages, Antibodies and De-monstration", *History and Philosophy of the Life Sciences*, 30(2): 131-157.

Cameron G., 2007, Interview with Miguel Garcia-Sancho, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK.

Campbell-Kelly M., 2003, *From Airline Reservations to Sonic the Hedgehog: A History of the Software Industry*, Cambridge: MIT Press.

Campbell-Kelly M. and Aspray W., 1996, *Computer: A History of the Information Machine*, New York: Basic Books.

Castells M., 1996, *The Information Age: Economy Society and Culture*, Malden: Blackwell Publishing, three vols.

Ceruzzi P., 2000, *A History of Modern Computing*, Cambridge: MIT, Second Edition.

Chow-White P. and García-Sancho M., in press, "Bidirectional Shaping and Spaces of Convergence: Interactions between Biology and Computing from the First DNA Sequencers to Global Genome Databases", *Science, Technology and Human Values.*

Codd E.F., 1970, "A Relational Model of Data for Large Shared Data Banks", *Communications of the ACM*, 13: 377-387.

Cook-Deegan R., 1994, *The Gene Wars: Science, Politics and the Human Genome*, London: W.W. Norton and Company.
Creager A. and Gaudillière J.P., 1996, "Meanings in Search of Experiments and Vice-versa: the Invention of Allosteric Regulation in Paris and Berkeley, 1959-1968", *Historical Studies in the Physical and Biological Sciences*, 27: 1-89.
Crick F., 1973, "Project K: the Complete Solution of 'E. coli'", *Perspectives in Biology and Medicine*, 17: 67-70.
Date C.J., 1981 [1975], *An Introduction to Database Systems*, London: Addison Wesley.
de Chadarevian S., 1996, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s", *Journal of the History of Biology*, 29: 361-386.
de Chadarevian S., 2002, *Designs for Life: Molecular Biology after World War II*, Cambridge: Cambridge University Press.
de Chadarevian S., 2004, "Mapping the Worm's Genome: Tools, Networks, Patronage," in: Rheinberger H.J. and Gaudillière J.P. (eds), *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Genetics*, London / New York: Routledge, 95-110.
de Chadarevian S. and Hopwood N. (eds), 2004, *Models: The Third Dimension of Science*, Stanford: Stanford University Press.
Doyle R., 1997, *On Beyond Living*, Stanford: Stanford University Press.
Eck R. and Dayhoff M., 1966, *Atlas of Protein Sequence and Structure*, Maryland: National Biomedical Research Foundation.
Edgerton D., 1999, "From Innovation to Use: Ten Eclectic Theses on the Historiography of Technology", *History and Technology*, 16: 111-136. French version available at *Annales HSS*, vols. 4-5, 1998.
Emmeche C., 1991, "A Semiotical Reflection on Biology, Living Signs and Artificial Life", *Biology and Philosophy*, 6: 325-340.
Etxeberria A. and Umerez J., 2006, "Organismo y Organización en la Biología Teórica: ¿Vuelta al organicismo?", *Ludus Vitalis*, XIV(26) : 3-38.
Falk R., 2008, "Molecular Genetics: Increasing the Resolving Power of Genetic Analysis", *History and Philosophy of the Life Sciences*, 30(1): 43-52.
Fortun M., 2008, *Promising Genomics: Iceland and deCODE Genetics in a World of Speculation*, Berkeley: University of California Press.
Fox Keller E., 1995, *Refiguring Life: Changing Metaphors in 20th Century Biology*, New York: Columbia University Press.
Fox Keller E., 2000, *The Century of the Gene*, Cambridge: Harvard University Press.
Fujimura J., 1999, "The Practices of Producing Meaning in Bioinformatics," in: Fortun M. and Mendelsohn E. (eds), *The Practices of Human Genetics*, Dordrecht: Kluwert, 49-88.
Fujimura J. and Fortun M., 1996, "Constructing Knowledge Across Social Worlds: The Case of DNA Sequence Databases in Molecular Biology", in: Nader L. (ed.) *Naked Science: Anthropological Inquiry into Boundaries, Power, and Knowledge*, New York / London: Routledge, 160-173.
García-Sancho M., 2007a, "The Rise and Fall of the Idea of Genetic Information (1948-2006)", *Genomics, Society and Policy*, 2(3): 16-36.

García-Sancho M., 2007b, "Mapping and Sequencing Information: the Social Context for the Genomics Revolution", *Endeavour*, 31(1): 18-23.
García-Sancho M., 2008, *Sequencing as a Way of Work: A History of its Emergence and Mechanisation – From Proteins to DNA, 1945-2000*, PhD Dissertation, Centre for the History of Science, Imperial College, London.
García-Sancho M., 2009, "The Perception of an Information Society and the Emergence of the First Computerized Biological Databases," in: Matsumoto A. and Nakano M. (eds), *Human Genome. Features, Variations and Genetic Disorders*, New York: Nova Publishers, 257-276.
García-Sancho M., 2010, "A New Insight Into Sanger's Development of Sequencing: from Proteins to DNA", *Journal of the History of Biology*, 43(2): 265-323.
Gibbons S., Kaye J., Smart A., Heeney C. and Parker M., 2007, "Governing Genetic Databases: Challenges Facing Research Regulation and Practice", *Journal of Law and Society*, 34(2): 163-189.
Gilbert W., 1992, "A Vision of the Grail," in: Kelves D.J. and Hood L. (eds), *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, Cambridge: Harvard University Press, 83-97.
Griffiths P., 2001, "Genetic Information: A Metaphor in Search of a Theory", *Philosophy of Science*, 68: 394-412.
Griffiths P.E. and Stotz K., 2006, "Genes in the Postgenomic Era", *Theoretical Medicine and Bioethics, 27(6): 499-521.*
Hagen J., 1999, "Naturalists, Molecular Biology, and the Challenge of Molecular Evolution", *Journal of the History of Biology*, 32 : 321-341.
Hagen J., 2001, "The Introduction of Computers into Systematic Research in the United States During the 1960s", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 32 (2): 291-314.
Haigh T., 2001, "Inventing Information Systems: the Systems Men and the Computer, 1950-1968", *The Business History Review*, 75(1): 15-61.
Haigh T., 2006a, "A Veritable Bucket of Facts: Origins of the Data base Management System", *ACM SIGMOD Record*, 35(2): 33-49.
Haigh T., 2006b, "Remembering the Office of the Future: the Origins of Word Processing and Office Automation", *Annals of the History of Computing*, 28(4): 6-31.
Hamm G., 2007, Phone Interview with Miguel Garcia-Sancho.
Hamm G. and Stüber K., 1982, "The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library", *Nucleotide Sequence Data Library News*, 1: 2-8.
Hamm G. and Cameron G., 1986, "The EMBL Data Library", *Nucleic Acids Research*, 14(1): 5-9.
Haraway D., 1976, *Crystals, Fabrics and Fields. Metaphors of Organicism in Twentieth-Century Developmental Biology*, New Haven: Yale University Press.
Haraway D., 1997, *Modest Witness at Second Millennium: Female Man Meets Oncomouse*, New York / London: Routledge.
Harper P., 2008, *A Short History of Medical Genetics*, Oxford: Oxford University Press.
Harvey M. and McMeekin A., 2007, *Public or Private Economies of Knowledge? Turbulence in the Biological Sciences*, Cheltenham: Edward Elgar.

Hayles N. K., 1999, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, Chicago / London: University of Chicago Press.
Hilgartner S., 2004, "Making Maps and Making Social Order: Governing American Genome Centers, 1988-93," in: Rheinberger H.I. and Gaudillière J.P. (eds), *From Molecular Genetics to Genomics; The Mapping Cultures of Twentieth Century Genetics*, London / New York: Routledge: 113-128.
Hoffmeyer J., 1996, *Signs of Meaning in the Universe*, Indiana: Indiana University Press.
Hoffmeyer J. and Emmeche C., 1991, "Code-Duality and the Semiotics of Nature," in: Anderson M. and Merrell F. (eds), *On Semiotic Modelling*, New York: Mouton de Gruyter: 117-166.
Hofschneider P. and Murray K., 2001, "Combining Science and Business: From Recombinant DNA to Vaccines Against Hepatitis B Virus," in: Buckel P. (ed.), *Recombinant Protein Drugs*, Basel / Boston / Berlin: Birkhäuser, 43-64.
Hood L., 1992, "Biology and Medicine in the Twenty First Century," in: Kevles D.J. and Hood L. (eds), *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, Cambridge: Harvard University Press, 136-163.
Kay L., 1993, *The Molecular Vision of Life: Caltech, the Rockefeller Foundation and the Rise of the New Biology*, Oxford: Oxford University Press.
Kay L., 1995, "Who Wrote the Book of Life? Information and the Transformation of Molecular Biology, 1945-55", *Science in Context*, 8 : 609-634.
Kay L., 2000, *Who Wrote the Book of Life: A History of the Genetic Code*, Stanford: Stanford University Press.
Kennard O., 1998, "From Private Data to Public Knowledge," in: Butterworth E. (ed.), *The Impact of Electronic Publishing on the Academic Community: An International Workshop Organized by the Academia Europaea and the Wenner-Gren Foundation*, London: Portland Press: 159-166.
Kennard O., 2007, Interview with Miguel García-Sancho, Cambridge, UK.
Keeney M., 1986, *Biotechnology: The University-Industrial Complex,* New Haven: Yale University Press.
Kernighan B. and Plauger P., 1976, *Software Tools*, London: Addison-Wesley.
Kline R., 2006, "Cybernetics, Management Science and Technology Policy", *Technology and Culture*, 47: 513-535.
Krige J., 2002, "The Birth of EMBO and the Difficult Road to EMBL", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 33: 547-564.
Lenoir T., 1999, "Shaping Biomedicine as an Information Science," in: Bowden M.E., Hahn T.B. and Williams R.V. (eds), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, Medford: ASIS Monograph Series: 27-45.
Lenoir T. (ed.), 2002, "Makeover: Writing the Body into the Posthuman Technospace", *Configurations*, 10(2, Part One) and 10(3, Part Two).
Leonelli S., 2008, "Bio-ontologies as Tools for Integration in Biology", *Biological Theory*, 3: 7-11.
Leonelli S., 2010, "Documenting the Emergence of Bio-ontologies: or, Why Researching Bioinformatics Requires HPSSB", *History and Philosophy of the Life Sciences*, 32(1): 105-126.

Maynard Smith J., 2000, "The Concept of Information in Biology", *Philosophy of Science*, 67: 177-194.
McKusick V., 1966, *Mendelian Inheritance in Man*, Baltimore: Johns Hopkins University Press.
Mindell D., 2002, *Between Human and Machine: Feedback, Control and Computing before Cybernetics*, Baltimore: Johns Hopkins University Press.
Mols S., 2006, *Error-Mindedness and the Computerisation of Crystallography, 1912-1955*, PhD Dissertation, Centre for the History of Science, University of Manchester.
Moody G., 2004, *Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine and Business*, London: Wiley.
Morange M., 2008, "The Death of Molecular Biology?", *History and Philosophy of the Life Sciences*, 30(1) : 31-42.
Moss L., 2004, *What Genes Can't Do*, Cambridge: MIT.
Müller-Wille S., 2003, "Joining Lapland and the Topinambes in flourishing Holland: Center and Periphery in Linnaean Botany", *Science in Context*, 16(4): 461-488.
Müller-Wille S. and Rheinberger H.-J., 2007, *Heredity Produced: At the Crossroads of Biology, Politics and Culture, 1500-1870*, Cambridge: MIT Press.
Murray K., 2007, Interview with Miguel García-Sancho. University of Edinburgh, UK.
November J., 2004, "LINC: Biology's Revolutionary Little Computer", *Endeavour*, 28(3): 125-131.
November J., 2006, *Digitizing Life: The Introduction of Computers to Biology and Medicine*, PhD Dissertation, Department of History, Princeton University.
Olazaran M., 1996, "A Sociological Study of the Official History of the Perceptrons Controversy", *Social Studies of Science*, 26(3): 611-659.
O'Malley M., 2007, "Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case", *History and Philosophy of the Life Sciences*, 29: 337-360.
Penders B., Horstman K., Vos R., 2008, "Walking the Line between Lab and Computation: the 'Moist' Zone", *BioScience*, 58(8): 747-755.
Pickstone J., 2000, *Ways of Knowing: A New History of Science, Technology and Medicine*, Chicago: University of Chicago Press.
Pickstone J., 2007, "Working Knowledges Before and After circa 1800: Practices and Disciplines in the History of Science, Technology, and Medicine", *Isis*, 98: 489-516.
Powell A., O'Malley M., Müller-Wille S., Calvert J., Dupré J., 2007, "Disciplinary Baptisms: A Comparison of the Naming Stories of Genetics, Molecular Biology, Genomics and Systems Biology", *History and Philosophy of the Life Sciences*, 29: 5-32.
Ramillon V., 2007, *Le deux génomiques. Mobiliser, organiser, produire: du séquençage à la mesure de l'expression des gènes*, PhD dissertation, École des Hautes Études en Sciences Sociales.
Rheinberger H.J., 2006, "The Notions of Regulation, Information, and Language in the Writings of François Jacob", *Biological Theory*, 1(3): 261-267.

Rosemberg D., 2003, "Early Modern Information Overload", special issue of the *Journal of the History of Ideas*, 64(1).
Sarkar S., 1996a, "Decoding Coding: Information and DNA", *Bioscience*, 46: 857-863. Reprinted in Sarkar S., 2005, *Molecular Models of Life: Philosophical Papers on Molecular Biology*, Cambridge: MIT Press, 183-204.
Sarkar S., 1996b, "Biological Information: A Skeptical Look at Some Central Dogmas in Molecular Biology," in: Sarkar S. (ed.), *The Philosophy and History of Molecular Biology: New Perspectives* (Dordrecht: Kluwer): 187-233. Reprinted in: Sarkar S., 2005, *Molecular Models of Life: Philosophical Papers on Molecular Biology,* Cambridge: MIT Press: 205-260.
Sarkar S., 1998, *Genetics and Reductionism*, Cambridge: Cambridge University Press.
Sarkar S., 2005, "How Genes Encode Information for Phenotypic Traits," in: Sarkar S., 2005, *Molecular Models of Life: Philosophical Papers on Molecular Biology*, Cambridge: MIT Press, 261-283.
Shannon C., 1948, "The Mathematical Theory of Communication", *Bell System Technical Journal*", 28(4): 656-715.
Secord J.A. and Jardine N. (eds), 1996, *Cultures of Natural History*, Cambridge: Cambridge University Press.
Segal J., 2003, *Le zéro et le un: histoire de la notion scientifique d'information au 20e siècle*, Paris: Syllepse.
Smith P.R., 1974, "Reservations on Project K", *Perspectives in Biology and Medicine*, 18: 21-23.
Smith T., 1990, "The History of the Genetic Sequence Databases", *Genomics*, 6: 701-707.
Star S.L. and Griesemer J., 1989, "Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39", *Social Studies of Science*, 19(3): 387-420.
Stegmann U., 2005, "Genetic Information As Instructional Content", *Philosophy of Science*, 72: 425-443.
Stotz K. and Griffiths P., 2004, "Genes: Philosophical Analyses Put to the Test", *History and Philosophy of the Life Sciences,* 26(1): 5-28.
Strasser B., 2003, "The Transformation of the Biological Sciences in Post-war Europe", *EMBO Reports*, 4(6): 540-543.
Strasser B., 2006, "Collecting and Experimenting: The Moral Economies of Biological Research, 1960s-1980s," in: de Chadarevian S. and Rheinberger H.J. (eds), *History and Epistemology of Molecular Biology and Beyond: Problems and Perspectives*, Berlin: Max Planck Institute for the History of Science, Preprint number 310, 105-123.
Strasser B., 2008, "Genbank: Natural History in the 21st Century?", *Science*, 322: 537-538.
Strasser B., 2010, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's 'Atlas of Protein Sequence and Structure', 1954-1965", *Journal of the History of Biology*, 43(4): 623-660.
Suárez Díaz E., 2007, "The Rhetoric of Informational Molecules: Authority and Promises in the Early Study of Molecular Evolution", *Science in Context*, 20(4): 649-677.

Suárez Díaz E., 2009, "Molecular Evolution: Concepts and the Origin of Disciplines", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 40: 43-53.

Suárez Díaz E., 2010, "Making Room for New Faces: Evolution, Genomics and the Growth of Bioinformatics", *History and Philosophy of the Life Sciences*, 32: 65-90.

Suárez Díaz E. (ed.), 2007, "Science and Representation: A Historical and Philosophical Approach", *History and Philosophy of the Life Sciences*, 29(2).

Suárez Díaz E. and Anaya Muñoz V., 2008, "History, Objectivity and the Construction of Molecular Phylogenies", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 39 : 451-468.

Tauber A. and Sarkar S., 1992, "The Human Genome Project: Has Blind Reductionism Gone So Far?", *Perspectives in Biology and Medicine*, 35(2): 220-235.

Taylor P.J., 2008, "The Under-recognized Implications of Heterogeneity: Opportunities for Fresh Views on Scientific, Philosophical and Social Debates about Heritability", *History and Philosophy of the Life Sciences*, 30(3-4): 431-456.

Webster F., 1997, "Is This the Information Age? Towards a Critique of Manuel Castells", *City*, 8: 71-84.

Wilson M., 1997, *The Difference Between God and Larry Ellison: Inside Oracle Corporation*, New York: William Morrow and Company.

Zweiger G., 2001, *Transducing the Genome: Information, Anarchy and Revolution in the Biomedical Sciences*, New York: McGraw Hill.